Document de travail
Working paper

Les comparaisons internationales d'état de santé subjectif sont-elles pertinentes ? Une évaluation par la méthode des vignettes-étalons

Salim Lardjane (CREST-ENSAI)
Paul Dourgnon (IRDES)

DT n° 2

Février 2007

Les comparaisons internationales d'état de santé subjectif sont-elles pertinentes ?

Une évaluation par la méthode des vignettes-étalons

Salim Lardjane (CREST-ENSAI, Laboratoire de Statistique d'Enquêtes)
Paul Dourgnon (IRDES Institut de Recherche et Documentation en Economie de la Santé)

Résumé

Lorsque les modalités de réponse à une question de santé subjective sont utilisées différemment par différents individus, on dit que celles-ci sont affectées d'un effet DIF. Nous présentons une méthode non-paramétrique de détection et de correction de tels effets dans les auto-évaluations de santé subjective. Nous illustrons cette méthode par la mise en évidence d'un effet DIF potentiel dans le cas de l'auto-évaluation du niveau de douleur physique ressenti par des personnes âgées issues de différents pays européens et montrons comment le prendre en compte dans l'analyse statistique.

1. Introduction.

La mesure de la santé des individus est essentielle pour suivre l'évolution de l'état de santé des populations et évaluer les interventions de santé dans les politiques publiques. Toutefois, il n'est pas toujours possible de mesurer les différentes dimensions de la santé étudiées à l'aide de procédés de mesures physiques ou d'avis de médecins, en raison de coûts trop élevés ou d'impossibilités pratiques. On a alors recours à des mesures subjectives de santé, typiquement obtenues à l'aide de questionnaires d'auto-évaluation (Falissard, 2001). Ces derniers peuvent être construit, lors de la conception de l'étude, en sélectionnant une ou plusieurs dimensions de santé, par exemple dans la classification ICIDH-2 mise au point par l'OMS (WHO 1980, 1999), puis en associant à chacune d'entre-elles une ou plusieurs questions d'auto-évaluation. En répondant à une question d'auto-évaluation, un individu évalue lui-même son état de santé, pour une dimension de santé donnée. Le questionnaire utilisé pour l'enquête SHARE 2004 (encadré 1) illustre bien cette approche.

< INSERER ENCADRE 1 >

Dans le questionnaire SHARE 2004, une question d'auto-évaluation est associée à chaque dimension de santé retenue et cinq modalités de réponses sont proposées pour chaque question d'auto-évaluation. Les dimensions de santé considérées et les questions d'auto-évaluation correspondantes sont directement reprises des questionnaires conçus par l'OMS pour l'Enquête Mondiale sur la Santé 2002-2003. A titre d'exemple, la question d'auto-évaluation associée à la dimension de santé *douleur physique*, que nous utilisons dans ce travail, est fournie dans l'encadré 2.

< INSERER ENCADRE 2 >

2. L'effet DIF en enquêtes de santé : présentation et modalités d'étude.

Supposons que les répondants partagent les mêmes attentes ou une même norme de santé pour une dimension de santé donnée. Dans ce cas, on a de bonnes raisons d'avancer que deux individus ayant répondu par une même modalité à une question d'auto-évaluation associée ont des niveaux de santé subjective relativement proches. Dans une telle situation, un individu ayant répondu « *léger* » à la question d'auto-évaluation précédente sur la douleur physique (encadré 2) se perçoit en meilleure santé qu'un individu ayant répondu « *moyen* » à cette même question.

Supposons à présent que deux répondants aient des attentes ou des normes très différentes pour une même dimension de santé. Dans ce cas, les réponses obtenues ne permettent pas nécessairement de comparer l'état de santé subjectif des individus. Ainsi, une personne atteinte de troubles respiratoires peut utiliser la modalité « *léger* » en réponse à une question d'auto-évaluation sur sa mobilité pour signifier qu'elle ne pense pas parvenir à effectuer 1 km sans s'essouffler, alors qu'un athlète professionnel peut utiliser cette même modalité « *léger* » pour signifier qu'il ne pense pas parvenir à effectuer ses 20 km de course hebdomadaires habituels. Dans une telle situation, certes extrême, les modalités de réponse sélectionnées ne permettent pas de comparer la mobilité subjective des deux individus.

Cet exemple illustre le fait que les modalités de réponse à la question d'auto-évaluation peuvent être utilisées différemment par les différents répondants et donc qu'elles peuvent « *fonctionner* » différemment pour les différents répondants. De façon plus formelle, on dit alors que les modalités de réponses – les *items* en psychométrie - sont affectée par un *biais d'item* ou, de façon plus neutre, d'un *effet de type DIF* (ang. Differential Item Functioning – Nunnaly & Bernstein 1994, Falissard 2001).

La présence d'un effet DIF affectant les réponses à une question d'auto-évaluation est toujours une possibilité théorique mais que, *sans plus d'information, cette possibilité doit être écartée* pour des raisons tant éthiques que scientifiques (principe de parcimonie). Il est donc fondé, sans plus d'information, de supposer *qu'il n'y a pas* d'effet DIF et donc pas de problème de comparabilité des réponses obtenues à une question d'auto-évaluation.

Diverses études comparatives de morbidité et de santé (Murray & Chen 1992, Salomon & al. 2004, Sen 2002, entre autres) ont toutefois montré que de tels problèmes de comparabilité des réponses étaient potentiellement présents dans les enquêtes sur la santé subjective, notamment lorsqu'on souhaitait comparer des échantillons issus de pays ou régions différents pour une ou plusieurs dimensions de santé, ce qui est la situation typique des enquêtes effectuées sous l'égide d'organisations internationales, telles l'ONU, l'OMS ou l'UE. Ces travaux ont également montré que l'occultation des problèmes de comparabilité des réponses pouvait conduire à des *contresens graves* sur l'état de santé des populations, notamment des plus défavorisées économiquement. Au final, les résultats de ces études, confrontés aux considérations précédentes sur l'effet DIF, plaident en faveur du recueil *d'informations supplémentaires* en complément des questions d'auto-évaluation lors d'enquêtes de santé subjective, afin de pouvoir détecter d'éventuels problèmes de comparabilité des réponses.

La méthode présentée dans ce travail est une illustration de cette approche. En exploitant des questions complémentaires associées à la question d'auto-évaluation sur la douleur physique dans le questionnaire SHARE 2004, nous mettons en évidence un effet DIF potentiel et montrons comment le prendre en compte lors de comparaisons entre échantillons. L'approche suivie paraît d'autant plus naturelle pour la douleur physique qu'il s'agit là d'une dimension de santé fondamentalement subjective. Elle peut également être utilisée pour d'autres dimensions de santé subjective et pour obtenir des hypothèses de nature comparative sur la

santé subjective au sein de sous-populations d'une population donnée, définies par des caractéristiques démographiques ou socio-économiques par exemple.

3. Vignettes d'ancrage (vignettes-étalons).

Ce travail se situe dans la continuité de méthodes récentes d'analyse et de traitement des effets DIF dans les enquêtes de santé, développées autour de l'OMS et du département de Sciences Politiques de l'Université de Harvard depuis le début des années 2000. Ces méthodes reposent fondamentalement sur l'inclusion d'éléments supplémentaires, appelés *vignettes d'ancrage* (ang. *anchoring vignettes*, *anchors*), que nous proposons d'appeler également *vignettes-étalons*, dans les questionnaires de santé subjective. C'est l'intérêt récent pour ces méthodes et leur utilité potentielle dans une optique de comparaisons internationales qui explique que de telles vignettes aient été incluses dans divers questionnaires de santé de l'OMS et dans le questionnaire SHARE 2004.
De façon générale, le terme *vignette* est utilisé en sciences sociales pour désigner toute *description d'une situation fictive* soumise à un individu, à laquelle on lui demande de réagir ou qu'on lui demande d'évaluer. L'utilisation de celles-ci remonte aux années 1950, avec des travaux en anthropologie, psychologie et sociologie, mais également en droit et en sciences de l'éducation (réf. in Alexander & Becker 1978, Salomon & al. 2004). Des vignettes ont, par ailleurs, été récemment utilisées en médecine, notamment en lien avec la formation des médecins et des infirmiers (réf. in Salomon & al. 2004).
La méthode des vignettes-étalons se distingue cependant des utilisations précédentes des vignettes par diverses spécificités (Salomon & al. 2004). D'une part, les méthodes utilisées historiquement utilisaient des vignettes obtenues à partir d'un canevas unique en faisant varier diverses caractéristiques de la situation décrite et d'autre part, les différentes vignettes ainsi obtenues étaient réparties aléatoirement entre les personnes enquêtées (Alexander & Becker 1978). Dans la cadre de la méthode des vignettes étalons, par contre, les vignettes ne dont pas nécessairement obtenues à partir d'un même canevas et chaque vignette doit être soumise à l'ensemble des individus enquêtés. De plus, dans l'optique d'une correction d'un effet DIF, des questions sensiblement identiques et des modalités de réponse identiques doivent être utilisées pour l'auto-évaluation et l'évaluation des vignettes.
Ainsi, on appellera spécifiquement *vignette d'ancrage ou vignettes-étalons* associée à une question d'auto-évaluation toute vignette soumise aux individus enquêtés, portant sur la même dimension de santé que la question d'auto-évaluation et qu'on demande aux individus d'évaluer sur l'échelle utilisée pour la question d'auto-évaluation, en réponse à une question aussi proche que possible de celle utilisée pour l'auto-évaluation. A titre d'illustration, les différentes vignettes-étalons associées à la question d'auto-évaluation portant sur la douleur physique dans le questionnaire SHARE 2004 sont présentées dans l'encadré 3.

< INSERER ENCADRE 3 >

Dans l'optique d'une utilisation pour détecter et corriger un effet DIF, les vignettes étalons et le protocole d'enquête doivent être conçus de façon à ce que les hypothèses suivantes puissent être considérés comme raisonnables.

H1. *Hypothèse d'équivalence des vignettes* : Tous les répondants comprennent de la même façon la situation décrite dans chacune des vignettes.

Cette hypothèse signifie que chaque vignette suffit à décrire de façon satisfaisante une situation qui est la même pour tous les individus enquêtés. Dans le cas de la douleur physique, cela revient à supposer que les individus évaluent bien une même situation en terme de douleur lors de l'évaluation d'une vignette. On mesure l'importance d'une bonne conception des vignettes et d'une traduction de qualité pour que cette hypothèse soit approximativement vérifiée dans le cadre d'enquêtes plurilingues.

H2. *Hypothèse d'équivalence des modalités de réponses* : Chaque répondant utilise les modalités de réponse de la même façon lors de l'auto-évaluation et lors de l'évaluation des différentes vignettes.

D'un point de vue pratique, H2 autorise à *comparer* formellement, à partir des réponses obtenues, l'auto-évaluation par un individu de son état de santé à son évaluation des différentes vignettes, et conditionne donc pour une large part l'utilisation des vignettes-étalons pour la mise en évidence et la correction d'un effet DIF potentiel. Une bonne formulation de la question et des modalités de réponses et un bon travail d'explication éventuel sont essentiels pour que cette hypothèse soit approximativement vérifiée. Notons que celle-ci *n'implique pas* que les *différents* répondants utilisent les modalités de réponse de la même façon, auquel cas il n'y aurait pas d'effet DIF.

Divers auteurs augmentent l'hypothèse d'équivalence des vignettes d'une hypothèses portant sur l'unidimensionnalité des évaluations ou la font implicitement – nous avons choisi de l'énoncer séparément en raison de l'usage qui en sera fait et de l'importance de l'hypothèse d'équivalence des vignettes, telle qu'énoncée ci-dessus, pour les comparaisons internationales.

H3. *Hypothèse d'unidimensionnalité* : L'auto-évaluation et l'évaluation des différentes vignettes se rapportent à une même dimension de santé pour tous les répondants.

Cette hypothèse permet d'une part, de considérer que les comparaisons formelles déduites des réponses d'un individu portent effectivement sur une même dimension de santé et d'autre part, de considérer qu'on a bien affaire à une même dimension de santé pour les différents répondants.

Concrètement, pour la douleur physique dans l'enquête SHARE 2004 (encadrés 2 et 3), les hypothèses précédentes sont vérifiée dans la une situation où chaque individu enquêté se « met à la place » de l'individu décrit dans chacune des vignettes et évalue la douleur physique correspondante, et uniquement la douleur physique correspondante, comme si c'était lui-même qui la subissait. L'expérience de la douleur physique étant fondamentalement subjective, l'aspect le plus critique est ici sans doute le premier.

Les hypothèses précédentes suffisent à justifier l'utilisation des vignettes de référence pour la *détection* d'un effet DIF. On doit toutefois leur adjoindre une hypothèse supplémentaire dans l'optique d'une *correction* de l'effet DIF.

H4. *Hypothèse d'ordre privilégié* : Les vignettes sont strictement ordonnées pour la dimension de santé considérée, selon $V1 < V2 < \ldots < Vk$ par ordre de gravité strictement croissant.

Dans le cadre d'une étude donnée, k, V1, …, Vk sont explicites. Par exemple, pour la douleur physique dans l'enquête SHARE 2004, nous ferons l'hypothèse que V1 < V2 < V3 (voir l'encadré 2 pour les définitions des vignettes V1 à V3). Nous verrons que cet ordre, qui peut sembler assez naturel, n'est pas partagé par tous les répondants.

L'ordre strict mentionné dans l'hypothèse H4 sera appelé *ordre privilégié*. Pour éviter une trop grande instabilité des résultats obtenus, il est souhaitable qu'il y ait consensus des individus enquêtés, des analystes, d'experts ou des utilisateurs de l'enquête sur l'ordre privilégié des vignettes ; en l'absence d'un tel consensus, on doit faire face à de redoutables problèmes d'agrégation des préférences, classiques en sciences politiques et économiques. *Cet aspect, peu abordé dans la littérature sur les vignettes étalons, nous semble être une limitation fondamentale de la méthodologie proposée, qui peut nuire significativement à la crédibilité de l'analyse, en particulier dès lors que le nombre de vignettes est trop élevé.* En effet, nous verrons que la procédure proposée pour corriger l'effet DIF repose de façon essentielle sur le fait que les évaluations des vignettes par les répondants soient consistantes avec l'hypothèse H4, ce qui requiert en particulier que des niveaux de santé suffisamment distincts soient associés aux différentes vignettes. Il y a donc un arbitrage à effectuer, lors de la conception du questionnaire, entre le nombre de vignettes à inclure et une différentiation suffisante de celles-ci.

Enfin, notons que les hypothèses ci-dessus décrivent des situations idéales ; on peut donc difficilement s'attendre à ce qu'elles soient vérifiés exactement dans la pratique. Une validité approximative est toutefois souvent suffisante et divers arguments en faveur de celle-ci sont fournis par Salomon & al. (2002), King & al. (2004, 2007), Salomon & al. (2000). On mesure l'importance de la conception des vignettes et des modalités d'enquête et la nécessité de *tester* les vignettes avant toute utilisation à grande échelle ; à cet égard, diverses recommandations pratiques sont fournies par Salomon & al. (2000). De plus, la plupart des recommandations données dans la littérature pour la conception d'instruments de mesure subjectives restent naturellement valables ici (Falissard 2001, par exemple). On verra par ailleurs comment étudier la pertinence des hypothèses de travail à l'aide d'outils statistiques et, de façon connexe, comment prendre en compte les répondants dont les évaluations des vignettes ne sont pas consistantes avec l'hypothèse H4, lors de la correction d'un effet DIF potentiel.

4. Détection d'un effet DIF potentiel.

Une première utilité des vignettes-étalons est qu'elles peuvent êtres utilisées, sous les hypothèses H1, H2, H3, pour *détecter* un effet DIF potentiel entre individus ou entre groupes d'individus.

Pour mettre en évidence un effet DIF potentiel et donc un problème de comparabilité des auto-évaluations entre deux répondants, il suffit de comparer les modalités de réponse utilisées pour l'évaluation des différentes vignettes. En effet, *si deux individus utilisent des modalités de réponse différentes lors de l'évaluation d'une même vignette*, les hypothèses d'équivalence des vignettes et d'équivalence des modalités de réponse impliquent qu'ils utilisent différemment les modalités de réponse lors de l'auto-évaluation ; par conséquent, en dépit de l'hypothèse d'unidimensionnalité, *leurs santés subjectives ne peuvent être directement comparées à partir de leurs réponses à la question d'auto-évaluation*. Cette conclusion étant obtenue sous les hypothèses H1 et H2, nous préférons parler dans ce cas d'effet DIF *potentiel*.

De façon analogue, pour mettre en évidence un effet DIF potentiel entre (sous-)populations, il suffit de comparer les fréquence relatives d'utilisation des différentes modalités de réponses utilisées à l'issue de l'évaluation des différentes vignettes dans chaque (sous-)population.

Les graphiques 1 à 3, résumant, par origine de l'échantillon, l'usage des différentes modalités de réponse pour chacune des vignettes associées à la question d'auto-évaluation portant sur la douleur physique dans le questionnaire SHARE 2004, permettent d'illustrer cette utilisation des vignettes étalons. *A notre sens, celle-ci suffit à elle seule à justifier l'inclusion de vignettes étalons dans des questionnaires comprenant des questions de santé subjective.*

<< INSERER GRAPHIQUE 1 >>

Le graphique 1 met en évidence, entre autres, une différence visuellement importante d'utilisation des modalités de réponse lors de l'évaluation de la vignette V1, entre l'échantillon suédois d'une part et les échantillons néerlandais, belge néerlandophone et grec d'autre part.

<< INSERER GRAPHIQUE 2 >>

Le graphique 2 révèle une différence visuellement importante d'utilisation des modalités de réponse lors de l'évaluation de la vignette V2, entre l'échantillon suédois d'une part et les autres échantillons d'autre part. Le graphique 3 ci-dessous met en évidence la même différence pour ce qui est de la vignette V3.

<< INSERER GRAPHIQUE 3 >>

Les remarques précédentes amènent à postuler l'existence d'un effet DIF entre échantillons et donc un problème de comparabilité des auto-évaluations obtenues, notamment entre l'échantillon suédois et les autres échantillons. La crédibilité de ce postulat repose essentiellement sur celle des hypothèses d'équivalence des vignettes et d'équivalence des modalités de réponse. L'hypothèse d'ordre privilégié n'est, quant à elle, pas utilisée à ce stade de l'analyse.

5. Correction d'un effet DIF

Une conséquence importante des hypothèses d'équivalence des vignettes, d'équivalence des modalités de réponse et d'unidimensionnalité est que le *positionnement de l'auto-évaluation d'un répondant par rapport à ses évaluations des vignettes associées* n'est pas affecté par l'effet DIF et peut donc être *déduit* de ses réponses. Ce positionnement peut alors être utilisé pour effectuer des comparaisons entre individus. En effet, un individu évaluant son niveau de santé avec une modalité supérieure (en gravité) à celle utilisée pour une vignette Vj donnée est en moins bonne santé subjective qu'un individu dont les évaluations des vignettes sont globalement dans le même ordre que celles de l'individu précédent et qui évalue son niveau de santé avec une modalité inférieure (en gravité) à celle utilisée pour cette même vignette Vj.

Dans la situation idéale où l'ordre sur les vignettes qui est déduit d'une comparaison de leurs évaluations est le même pour tous les répondants, il est possible de comparer l'état de santé de deux répondants quelconques. Lorsque c'est le cas, on peut naturellement adopter cet ordre

commun comme ordre privilégié dans l'hypothèse H4. Dans le cas plus général, et plus réaiste, où l'ordre privilégié n'est pas un ordre partagé par l'ensemble des répondants, on introduit la notion suivante.

Consistance : Les réponses d'un individu enquêté sont qualifiées de *consistantes* si l'évaluation qu'il fait des vignettes est consistante avec l'hypothèse H4 c'est-à-dire si elle respecte l'ordre privilégié.

Un individu dont les réponses sont *consistantes* et qui évalue son état de santé à l'aide d'une modalité supérieure (en gravité) à celle utilisée pour une vignette donnée, est en moins bonne santé subjective qu'un individu *dont les réponses sont consistantes* et qui évalue son propre état de santé à l'aide d'une modalité inférieure (en gravité) à celle utilisée pour cette même vignette, et ce *indépendamment de la population à laquelle il appartient*. Dans ce type de comparaisons, les vignettes sont utilisées comme des *étalons* ou *points d'ancrage* permettant de comparer les auto-évaluations. Afin de systématiser cette approche, King & al. (2004, 2007) proposent de définir, sous les hypothèses H1 à H4 et pour les répondants dont les réponses sont consistantes, une variable numérique ordinale C traduisant le positionnement de l'auto-évaluation par rapport aux évaluations des vignettes (encadré 4). La variable ainsi obtenue peut être interprétée comme une version de l'auto-évaluation *corrigée non-paramétriquement* de l'effet DIF (King & al. 2004, 2007).

< INSERER ENCADRE 4 >

La variable C peut être utilisée de façon très simple pour comparer l'état de santé de deux répondants dont les réponses sont consistantes. Par exemple, si $C(i) < C(j)$, on peut immédiatement affirmer que l'individu i est en meilleur santé subjective que l'individu j pour la dimension de santé considérée. Une limite évidente de cette approche est que C n'est définie que pour les répondants dont les réponses sont consistantes et, par conséquent, ne permet pas d'effectuer des comparaisons d'état de santé subjectif entre (sous-)populations comprenant des répondants dont les réponses ne sont pas consistantes. Afin de pallier ces inconvénients, King & Wand (2007) proposent une approche plus générale, que nous adaptons de façon originale dans ce travail. Il s'agit d'une solution qui n'est pas exempte de critique mais qui permet à tout le moins de mettre en évidence et de quantifier l'influence des problèmes de consistance sur l'issue de l'analyse.
L'idée développée par King & Wand (2007) consiste à identifier, pour chaque individu i dont les réponses ne sont pas consistantes, un *ensemble* G(i) de vignettes susceptibles d'être équivalentes à la situation de l'individu i pour la dimension de santé considérée. Une présentation détaillée et illustrée de la procédure d'obtention de G(i) est donnée dans King & Wand (2007) et une implémentation informatique de celle-ci est disponible dans le package R *anchors* développé par King & Wand (2005), que nous avons utilisé et adapté aux besoins de notre étude.
Nous utilisons les ensembles $G(\cdot)$ et la variable C pour définir, pour chaque répondant i, un *couple* de valeurs $(C(i,-) , C(i,+))$ décrivant le résultat de son auto-évaluation relativement aux différentes vignettes (encadré 5).

< INSERER ENCADRE 5 >

L'introduction des variables $C(\cdot,-)$ et $C(\cdot,+)$, bien qu'elle soit naturelle, est à notre connaissance, spécifique à notre travail. Nous proposons d'interpréter $C(i,-)$ comme une évaluation *optimiste* de la santé subjective de l'individu i et $C(i,+)$ comme une évaluation

pessimiste de la santé subjective de l'individu i. On dispose ainsi d'un *intervalle ordinal d'incertitude* sur l'état de santé subjectif de *chaque individu*, que ses réponses soient consistantes ou non.
Il est, à ce stade, important de noter que la construction effectuée et l'interprétation qui en est donnée est relative à l'ordre privilégié et qu'elle peut donc être critiquée sur cette base. Nous verrons toutefois que le fait d'obtenir des intervalles d'incertitude permet de *quantifier l'importance des problèmes de non-consistance lors de comparaisons entre (sous-)populations* et donc de mettre en évidence l'ampleur des écarts potentiels aux hypothèses de travail, ce qui permettra de relativiser certains résultats obtenus.

6. Utilisation des auto-évaluations corrigées de l'effet DIF pour des comparaisons entre (sous-)populations

Les variables $C(\cdot,-)$ et $C(\cdot,+)$ peuvent être utilisées pour obtenir, pour chaque population et chaque vignette Vj un *intervalle d'incertitude* pour la proportion d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette Vj et un *intervalle d'incertitude* pour la proportion d'individus évaluant leur niveau de santé inférieur ou équivalent à celui décrit dans la vignette Vj (encadré 6). Les intervalles obtenus pour diverses (sous-)populations peuvent alors être comparés à l'aide d'outils exploratoires, notamment graphiques.

< INSERER ENCADRE 6 >

Illustrons l'utilisation de ces intervalles d'incertitude par une comparaison des échantillons SHARE 2004 par origine, pour la dimension de santé douleur physique.

Le graphique 4 présente les intervalles d'incertitude obtenus dans les différents échantillons pour la proportion d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette V1 (encadré 2).

< INSERER GRAPHIQUE 4 >

L'échantillon suédois se distingue clairement des autres échantillons par une proportion plus élevée d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette V1 et par une incertitude très faible. L'échantillon néerlandais se distingue également de la plupart des autres échantillons, mais pas de l'échantillon italien, caractérisé par une incertitude très importante, donc par des problèmes de consistances important.

L'examen des proportions d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette V2 (encadré 2) pour la dimension douleur physique conduit à des conclusions analogues.

< INSERER GRAPHIQUE 5 >

On remarque que l'échantillon suédois se distingue par la proportion la plus élevée d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette V2 et par une incertitude très faible. On peut également avancer que la proportion d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette V2 est plus élevée

au sein de l'échantillon néerlandais qu'au sein de l'échantillon grec. Par contre, aucune des autres comparaisons entre échantillons ne met en évidence de différence visuellement significative, les différents intervalles d'incertitude se recouvrant partiellement les uns les autres. Dans le cas de l'échantillon italien, cette conclusion doit encore être relativisée par l'importance des problèmes de consistance.

Le graphique 6 présente les intervalles d'incertitude obtenus dans les différents échantillons pour la proportion d'individus évaluant leur niveau de santé inférieur ou équivalent à celui décrit dans la vignette V3 (encadré 2) pour la dimension douleur physique.

< INSERER GRAPHIQUE 6 >

La proportion d'individus évaluant leur niveau de santé inférieur ou équivalent à celui décrit dans la vignette V3 est moindre au sein de l'échantillon suédois qu'au sein des autres échantillons et l'incertitude est, là encore, très faible. D'autre part, les autre échantillons ne se distinguent pas clairement les uns des autres pour ce critère, même si, dans le cas de l'échantillon italien, cette conclusion doit être relativisée par l'importance des problèmes de consistance.

Les remarques précédentes amènent naturellement à postuler, au vu des données, que l'échantillon suédois est globalement en meilleure santé subjective que les autres échantillons en termes de douleur physique et qu'en dehors d'éventuels résultats partiels, concernant les échantillons néerlandais et grecs notamment, il n'y a aucune différence significative d'état de santé subjectif global entre les autres échantillons pour cette même dimension.

Ces résultats diffèrent significativement de ceux obtenus en se basant uniquement sur les auto-évaluations non corrigées de l'effet DIF et conduisent à relativiser les comparaisons basées sur celles-ci. A titre d'illustration, considérons les réponses à la question d'auto-évaluation obtenues pour échantillons suédois et néerlandais (tableau 1).

< INSERER TABLEAU 1 >

On voit que l'utilisation des auto-évaluations brutes amène naturellement à postuler que les échantillons suédois et néerlandais ne se distinguent pas vraiment en terme de santé subjective, voire qu'il y a davantage d'individus subjectivement en mauvaise santé, pour ce qui est de la douleur physique, au sein de l'échantillon suédois qu'au sein de l'échantillon néerlandais, ce qui contraste fortement avec les résultats de l'analyse précédente. Ainsi, l'analyse des auto-évaluations brutes, d'une part, et celles des auto-évaluations corrigées de l'effet DIF, d'autre part, amènent à formuler des hypothèses comparatives différentes sur l'état de santé subjectif global des échantillons suédois et néerlandais. Toutefois, l'utilisation *très* différentes des modalités de réponses lors de l'évaluation les vignettes au sein des échantillons suédois et néerlandais (graphiques 1 à 3) plaide fortement en faveur d'un effet DIF et d'une différence en termes d'attentes ou de norme pour la dimension de santé considérée entre l'échantillon suédois et néerlandais ; ce n'est bien entendu pas la seule explication possible, mais les résultats précédents incitent à aborder toute comparaison de la santé subjective des échantillons suédois et néerlandais à partir des auto-évaluations non corrigées avec circonspection.

7. Conclusion.

L'inclusion de vignettes étalons dans des questionnaires de santé subjective soumis à diverses (sous-)populations permet de détecter d'éventuels effets DIF pour des questions de santé subjective et donc de possibles différences d'attentes ou de normes de santé entre (sous-)populations. En présence de tels effets, les auto-évaluations ne peuvent être utilisées directement pour comparer l'état de santé subjectif global des (sous-)populations considérées et doivent être corrigées de l'effet DIF avant d'être utilisées dans une optique comparative. La méthode de correction proposée dans cet article repose sur des hypothèses de travail dont la validité dépend essentiellement de la qualité du questionnaire utilisé et des opérations de collecte. Elle permet de formuler des hypothèses comparatives sur l'état de santé subjectif global des (sous-)populations considérées relativement à des vignettes-étalons. Ces dernières se révèlent être non seulement un outil précieux pour l'étude de la pertinence des comparaisons internationales de santé subjective à partir d'auto-évaluations, mais également un outil de recherche et d'exploration intéressant.

Bibliographie

1. C. S. Alexander & H. J. Becker (1978). *The Use of Vignettes in Survey Research.* The Public Opinion Quarterly, Vol. 42, No.1, pp. 93-104.
2. A. Börsch-Supan (Coordinator), Hendrik Jürges (2005). *The Survey of Health, Aging, and Retirement in Europe – Methodology*. Mannheim Research Institute for the Economics of Aging, Mannheim.
3. B. Falissard (2001). *Mesurer la subjectivité en Santé*. Masson, Paris.
4. G. King, C. J. L. Murray, J. A. Salomon & A. Tandon (2004). *Enhancing the validity and cross-cultural comparability of survey research (final version).* American Political Science Review Vol. 98, No. 1.
5. G. King & J. Wand (2005). *Anchors: Software for Anchoring Vignette Data*. Documentation R, disponible à l'adresse http://GKing.Harvard.edu/vign/.
6. G. King & J. Wand (2007). *Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes*. Political Analysis – *à paraître*.
7. C. J. L. Murray & L. C. Chen (1992). *Uderstanding Morbidity Change*. Population and Development Reviews, Vol. 18, No. 3, pp. 481-503.
8. J. C. Nunnally & I. H. Bernstein (1994). *Psychometric Theory*. McGraw-Hill.
9. J. A. Salomon, C. J. L. Murray & A. Tandon (2002). *Using vignettes to improve cross-population comparability of health surveys : concepts, design, and evaluation techniques*. World Health Organization, GPE Discussion Paper No. 42.J.
10. A. Salomon, A. Tandon & C. J. L. Murray (2004). *Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes*. British Medical Journal, Vol. 328, pp. 258 & seq.
11. A. Sen (2002). *Health : Perception versus Observation*. British Medical Journal, 324, pp. 860-861.
12. World Health Organization (1980). *International Classification of Impairments, Disabilities, and Handicaps: a manual of classification relating to the consequences of disease*. World Health Organization, Geneva.
13. World Health Organization (1999). *ICIDH-2: International Classification of Functioning and Disability*. Beta-2 draft, short version. World Health Organization, Geneva.

ENCADRE 1 : ***Quelques éléments sur l'enquête SHARE 2004 et les données utilisées.***

L'enquête SHARE 2004 – *Survey of Health, Aging and Retirement in Europe* (Börsch-Supan & Jurges 2005) – a été réalisée dans 12 pays européens, donc en différentes langues. Cette enquête a été explicitement conçue dans le but de répondre à des problématiques de comparaisons internationales. Les divers questionnaires nationaux ont été obtenus à partir d'un questionnaire générique en anglais, qui a été traduit question par question. Dans ce travail, on limite le champ de l'étude aux individus enquêtés ayant répondu à la question d'auto-évaluation sur la douleur physique (encadré 2) et aux questions de vignettes associées (encadré 3) dans 8 pays : Allemagne, Belgique, Espagne, France, Grèce, Italie, Pays-Bas, Suède, avec une distinction entre Belgique francophone et Belgique néerlandophone. On dispose donc finalement de 9 échantillons.

ENCADRE 2 : ***Question d'auto-évaluation pour la dimension de santé douleur physique – questionnaire SHARE 2004.***

Dans l'ensemble, au cours des 30 derniers jours, quel niveau de douleurs physiques avez-vous ressenti ?

Aucun	Léger	Moyen	Grave	Extrême
☐$_1$	☐$_2$	☐$_3$	☐$_4$	☐$_5$

ENCADRE 3 : ***Vignettes associées à la dimension de santé douleur physique – questionnaire SHARE 2004***

Première vignette (vignette V1) associée au domaine de santé *douleur physique* dans le cadre de l'enquête SHARE 2004.

> *(V1) Paul a un mal de tête une fois par mois qui diminue après qu'il ait pris un cachet. Pendant qu'il a mal à la tête, il peut mener ses activités quotidiennes.*

Question associée :

> *En général, au cours des 30 derniers jours, quel niveau de douleurs physiques Paul a-t-il éprouvé ?*

Aucun	Léger	Moyen	Grave	Extrême
☐1	☐2	☐3	☐4	☐5

Deuxième vignette (vignette V2) associée au domaine de santé *douleur physique* dans le cadre de l'enquête SHARE 2004.

> *(V2) Henri a mal dans tout son bras droit et son poignet pendant sa journée de travail. Cela est partiellement atténué la soirée lorsqu'il ne travaille plus devant l'ordinateur.*

Question associée:

> *En général, au cours des 30 derniers jours, quel niveau de douleurs physiques Henri a-t-il éprouvé ?*

Aucun	Léger	Moyen	Grave	Extrême
☐1	☐2	☐3	☐4	☐5

Troisième vignette (vignette V3) associée à la dimension de santé *douleur physique* dans le cadre de l'enquête SHARE 2004.

> *(V3) Charles a mal aux genoux, aux coudes, aux poignets et aux doigts, et la douleur est presque continuellement présente. Bien que les médicaments aident, il ne se sent pas bien lorsqu'il se déplace, qu'il tient ou soulève quelque chose.*

Question associée :

> *En général, au cours des 30 derniers jours, quel niveau de douleurs physiques Charles a-t-il éprouvé ?*

Aucun	Léger	Moyen	Grave	Extrême
☐1	☐2	☐3	☐4	☐5

ENCADRE 4 : ***Définition d'une version des auto-évaluations corrigée de l'effet DIF pour les répondants dont les réponses sont consistantes – dimension de santé douleur physique, SHARE 2004.***

La réponse à la question d'auto-évaluation possède 5 modalités sur une échelle ordinale (encadré 1). A chaque individu i, on soumet 3 vignettes à évaluer sur la même échelle (encadré 2).
L'auto-évaluation de l'individu i est notée Y(i). L'évaluation de la vignette j par l'individu i est notée Z(i,j). Les vignettes associées à la question d'auto-évaluation sont numérotées de V1 à V3 par niveau de gravité strictement croissant, c'est-à-dire par niveau de santé strictement décroissant, selon l'ordre privilégié V1 < V2 < V3.

Avec cette convention, les réponses de l'individu i sont consistantes si et seulement si Z(i,1) < Z(i,2) < Z(i,3). Si c'est le cas, on pose

C(i) = 1 si Y(i) < Z(i,1)
C(i) = 2 si Y(i) = Z(i,1)
C(i) = 3 si Z(i,1) < Y(i) < Z(i,2)
C(i) = 4 si Y(i) = Z(i,2)
C(i) = 5 si Z(i,2) < Y(i) < Z(i,3)
C(i) = 6 si Y(i) = Z(i,3)
C(i) = 7 si Y(i) > Z(i,3)

Si C(i) est *pair*, cela signifie que la *même* modalité de réponse a été sélectionnée pour la question d'auto-évaluation et l'une des vignettes associées; le rang de cette dernière dans la série de vignettes ordonnée par gravité strictement croissante est C(i)/2. Si C(i) est *impair*, cela signifie que la modalité sélectionnée en réponse à la question d'auto-évaluation n'a été retenue pour aucune des vignettes; le niveau de santé de l'individu se positionne, en termes de rang dans la série des vignettes ordonnée par gravité strictement croissante, entre (C(i)-1) / 2 et (C(i)+1) / 2.

ENCADRE 5 : ***Définition d'un intervalle d'incertitude pour l'auto-évaluation corrigée de l'effet DIF– dimension de santé douleur physique, enquête SHARE 2004***

Pour un individu i dont les réponses ne sont pas consistantes, convenons de noter, r(i,-) le rang, dans la série des vignettes ordonnée selon l'ordre privilégié (hypothèse H4), de la vignette élément de G(i) correspondant au niveau de santé le moins grave et r(i,+) le rang, dans la série des vignettes ordonnée selon l'ordre privilégié, de la vignette élément de G(i) correspondant au niveau de santé le plus grave.

On pose alors $C(i,-) = 2\, r(i,-)$ et $C(i,+) = 2\, r(i,+)$.

C(i,-) correspond à la valeur de C qui serait associée à un individu i dont les réponses seraient consistantes et dont le niveau de santé pour la dimension considérée correspondrait à peu près à celui décrit dans la vignette de rang r(i,-).

C(i,+) correspond à la valeur de C qui serait associée à un individu i dont les réponses seraient consistantes et dont le niveau de santé pour la dimension considérée correspondrait à peu près à celui décrit dans la vignette de rang r(i,+).

Pour un individu i dont les réponses sont consistantes, on pose $C(i,-) = C(i,+) = C(i)$.

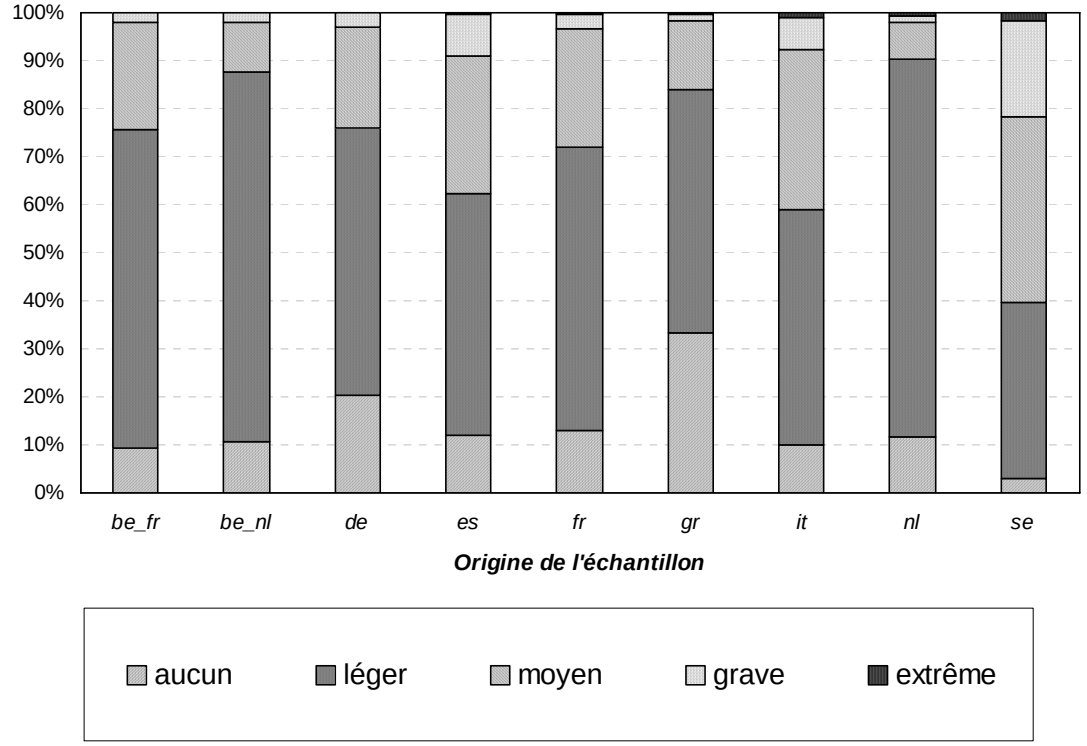
ENCADRE 6 : ***Définition des intervalles d'incertitudes (proportions)***.

Notons F(x,-) la proportion d'individus pour lesquels C(·,-) prend une valeur inférieure ou égale à x et F(x,+) la proportion d'individus pour lesquels C(·,+) prend une valeur inférieure ou égale à x dans l'une des (sous-)populations considérées. Alors, pour j = 1, 2, …, k où k désigne le nombre de vignettes, l'intervalle I(j,-) = [F(2j,+), F(2j,-)] est un *intervalle d'incertitude* pour la proportion d'individus dont le niveau de santé subjectif est supérieur ou équivalent à celui décrit dans la vignette Vj.

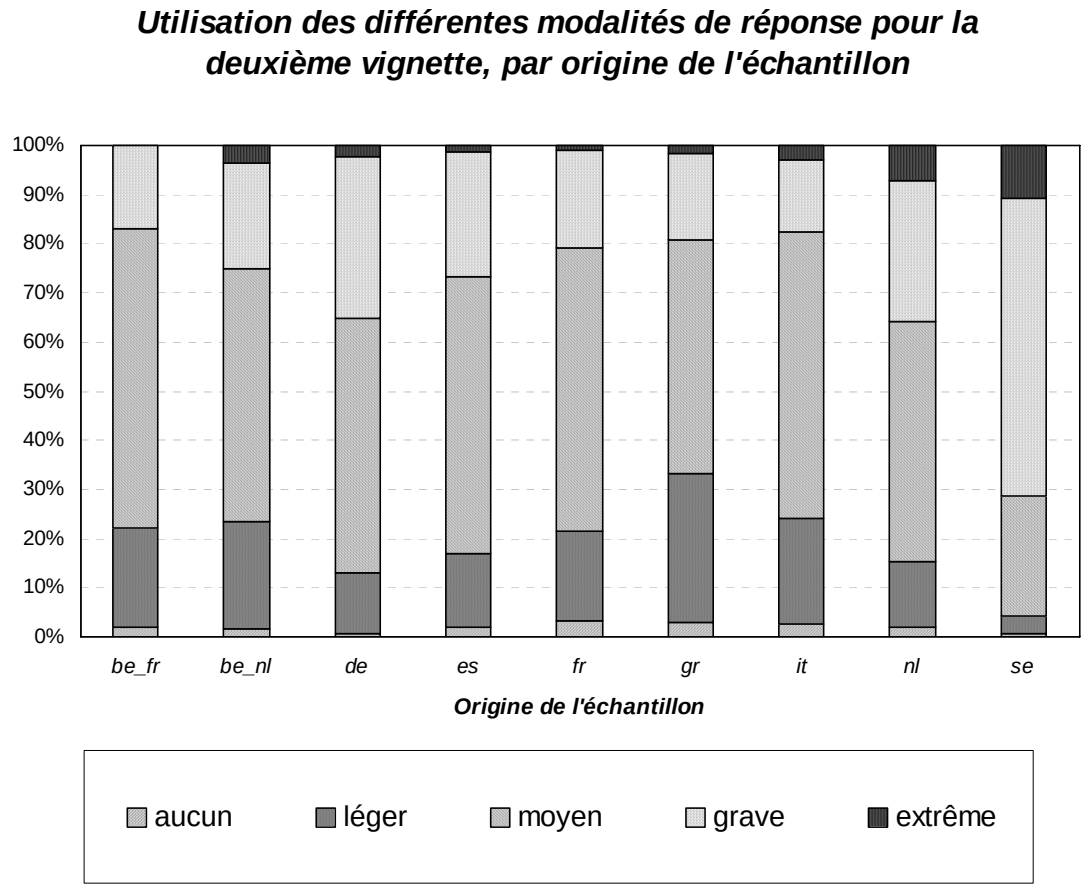
Notons G(x,-) la proportion d'individus pour lesquels C(·,-) prend une valeur supérieure ou égale à x et G(x,+) la proportion d'individus pour lesquels C(·,+) prend une valeur supérieure ou égale à x dans l'une des (sous-)populations considérées. Alors, pour j = 1, 2, …, k où k désigne le nombre de vignettes, l'intervalle I(j,+) = [G(2j,-), G(2j,+)] est un *intervalle d'incertitude* pour la proportion d'individus dont le niveau de santé subjectif est inférieur ou équivalent à celui décrit dans la vignette Vj.

GRAPHIQUE 1 : ***Utilisation des modalités de réponses à l'issue de l'évaluation de la vignette V1***

Utilisation des différentes modalités de réponse pour la première vignette, par origine de l'échantillon

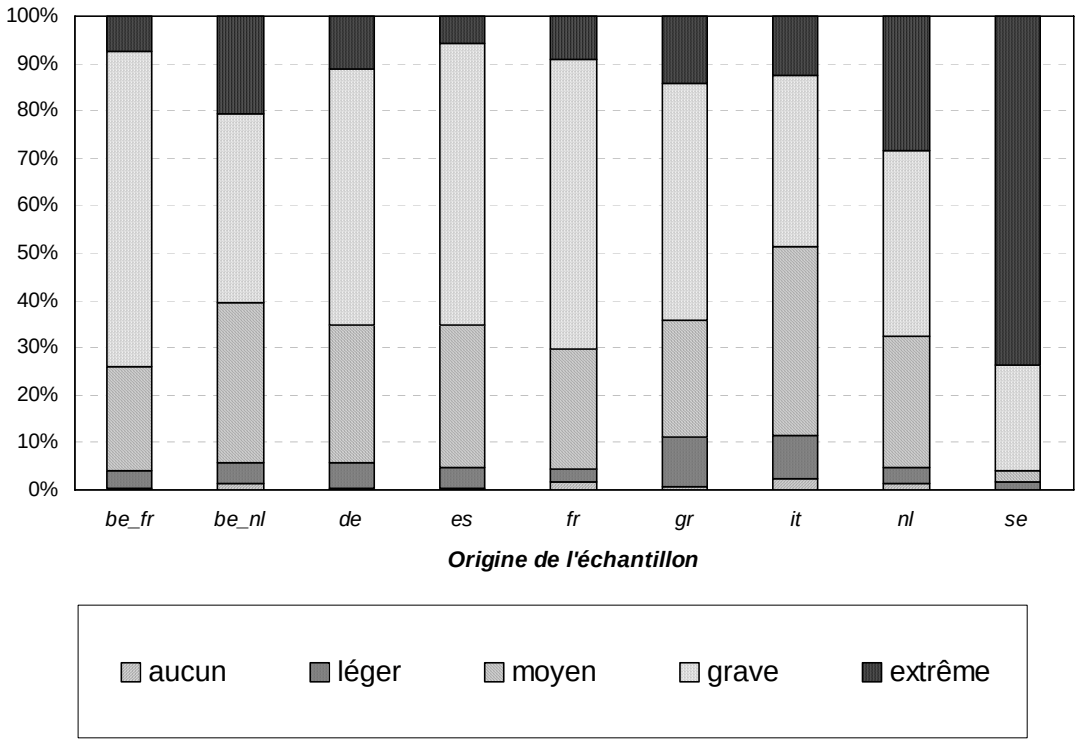


GRAPHIQUE 2 : ***Utilisation des modalités de réponses à l'issue de l'évaluation de la vignette V2***

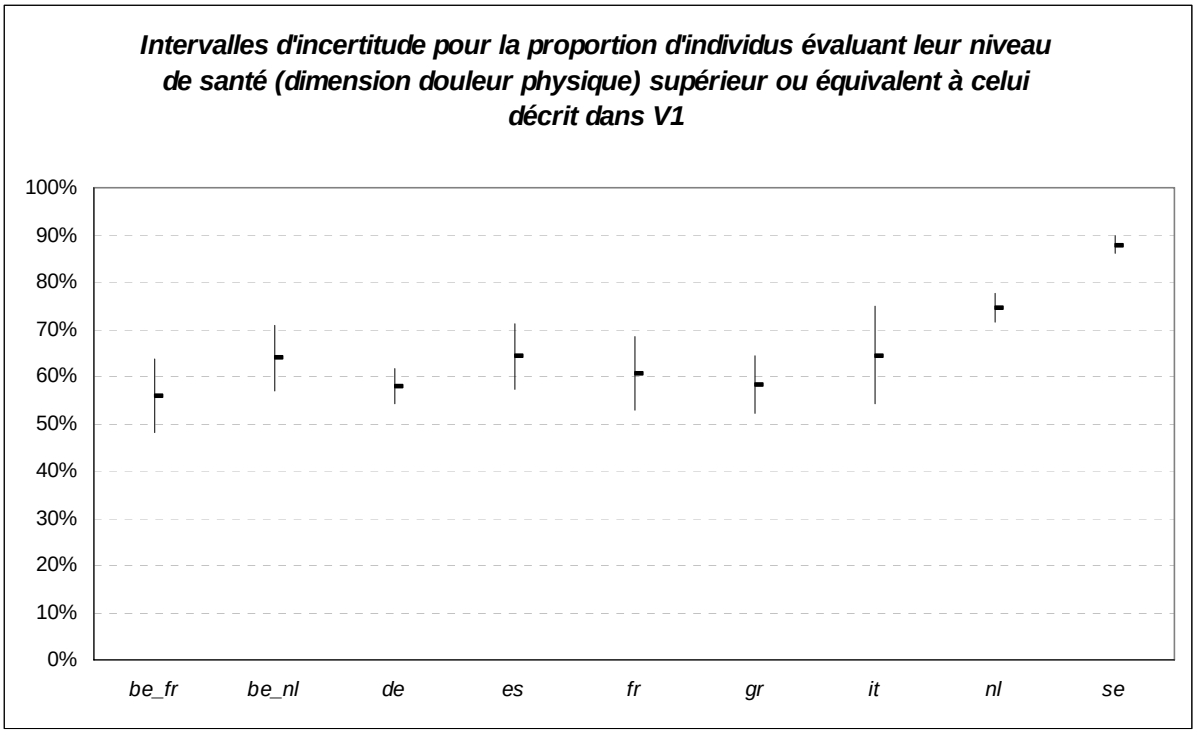


GRAPHIQUE 3 : ***Utilisation des modalités de réponses à l'issue de l'évaluation de la vignette V3***

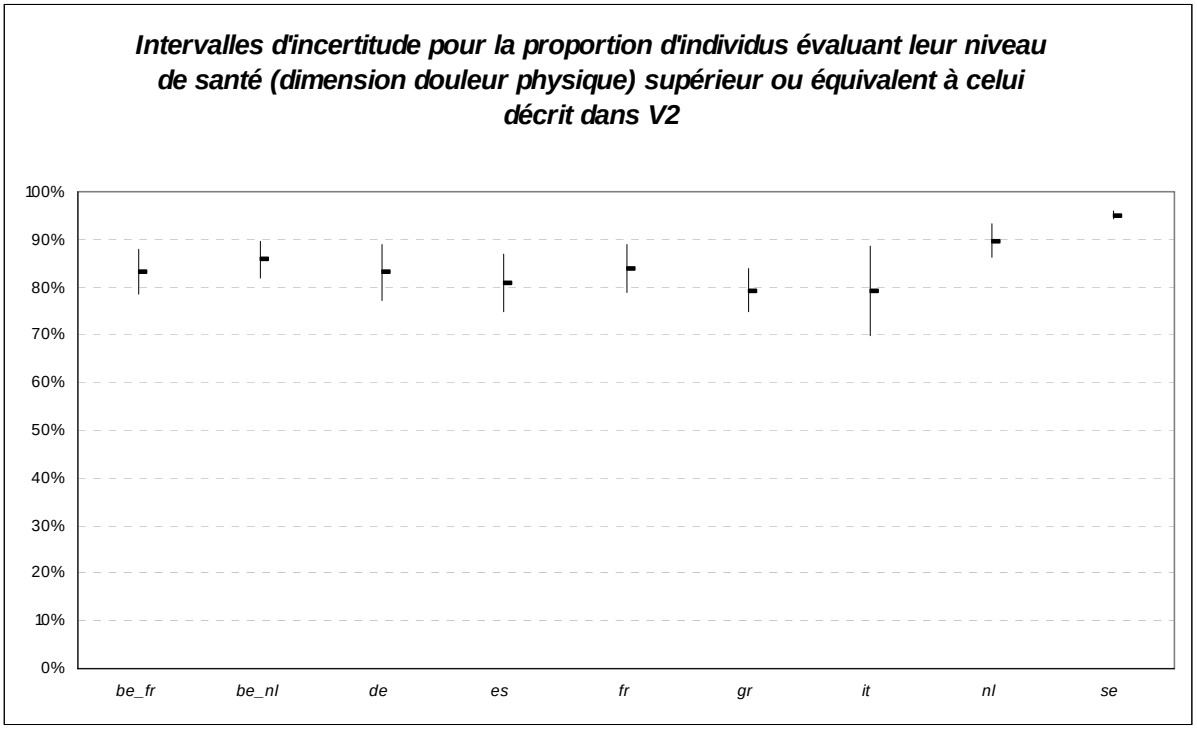
Utilisation des différentes modalités de réponse pour la troisième vignette, par origine de l'échantillon



GRAPHIQUE 4 : ***Intervalles d'incertitude pour la proportion d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette V1 pour la dimension de santé douleur physique, par origine de l'échantillon – données SHARE 2004***



GRAPHIQUE 5 : ***Intervalles d'incertitude pour la proportion d'individus évaluant leur niveau de santé supérieur ou équivalent à celui décrit dans la vignette V2 pour la dimension de santé douleur physique, par origine de l'échantillon – données SHARE 2004***



GRAPHIQUE 6 : ***Intervalles d'incertitude pour la proportion d'individus évaluant leur niveau de santé inférieur ou équivalent à celui décrit dans la vignette V3 pour la dimension de santé douleur physique, par origine de l'échantillon – données SHARE 2004.***

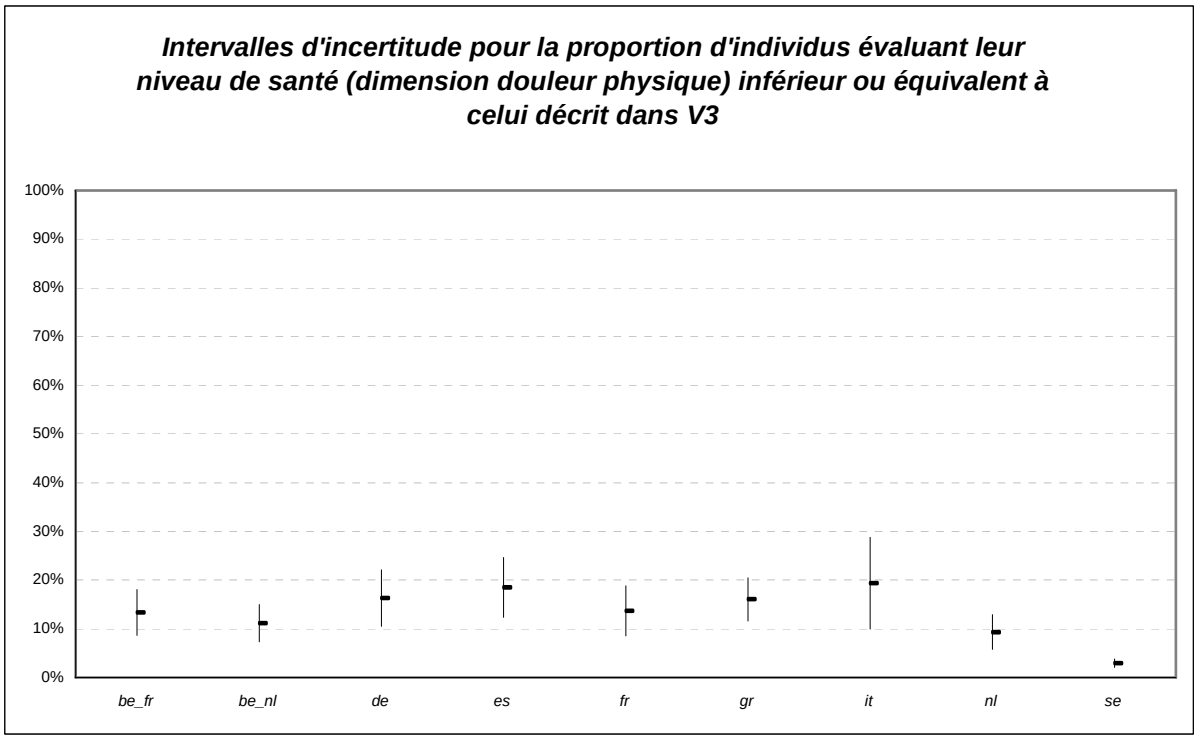


TABLEAU 1 : ***Fréquences cumulées de la réponses à la question d'auto-évaluation pour la dimension de santé douleur physique - répondants néerlandais et suédois, données SHARE 2004.***

	NL	SE
réponse = aucun	37,50%	52,55%
réponse ≤ léger	80,96 %	81,63 %
réponse ≤ moyen	94,81 %	92,09 %
réponse ≥ moyen	19,04%	18,37%
réponse ≥ grave	5,19 %	7,91 %
réponse = extrême	1,73%	1,02%